Journal of Machine Learning Research 12 (2011) 2389-2410 Submitted 3/10; Revised 10/10; Published 7/11

Universality, Characteristic Kernels and RKHS Embedding of Measures

Bharath K. Sriperumbudur BHARATH@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit
University College London
Alexandra House, 17 Queen Square
London WC1N 3AR, UK

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa
Tokyo 190-8562, Japan

Gert R. G. Lanckriet GERT@ECE.UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407, USA

Editor: John Shawe-Taylor

Abstract

Over the last few years, two different notions of positive definite (pd) kernels—universal and characteristic—have been developing in parallel in machine learning: universal kernels are proposed in the context of achieving the Bayes risk by kernel-based classification/regression algorithms while characteristic kernels are introduced in the context of distinguishing probability measures by embedding them into a reproducing kernel Hilbert space (RKHS). However, the relation between these two notions is not well understood. The main contribution of this paper is to clarify the relation between universal and characteristic kernels by presenting a unifying study relating them to RKHS embedding of measures, in addition to clarifying their relation to other common notions of strictly pd, conditionally strictly pd and *integrally strictly pd* kernels. For *radial* kernels on $\mathbb{R}^d$, all these notions are shown to be equivalent.

Keywords: kernel methods, characteristic kernels, Hilbert space embeddings, universal kernels, strictly positive definite kernels, integrally strictly positive definite kernels, conditionally strictly positive definite kernels, translation invariant kernels, radial kernels, binary classification, homogeneity testing

1. Introduction

Kernel methods have been popular in machine learning and pattern analysis for their superior performance on a wide spectrum of learning tasks. They are broadly established as an easy way to construct nonlinear algorithms from linear ones, by embedding data points into higher dimensional reproducing kernel Hilbert spaces (RKHSs) (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). In the regularization approach to learning (Evgeniou et al., 2000), it is well known that kernel-based algorithms (for classification/regression) generally invoke the *representer theorem* (Kimeldorf and Wahba, 1970; Schölkopf et al., 2001) and learn a function in a RKHS that has the

representation,

$$f := \sum_{j \in \mathbb{N}_n} c_j k(\cdot, x_j), \tag{1}$$

where $\mathbb{N}_n := \{1, 2, \ldots, n\}$, $k : X \times X \to \mathbb{R}$ is a symmetric positive definite (pd) kernel on some arbitrary space, X and $\{c_j : j \in \mathbb{N}_n\} \subset \mathbb{R}$ are parameters typically obtained from training data, $\{x_j : j \in \mathbb{N}_n\} \subset X$. As noted in Micchelli et al. (2006), one can ask whether the function, f in (1) approximates any real-valued target function arbitrarily *well* as the number of summands increases without bound. This is an important question to consider because if the answer is affirmative, then the kernel-based learning algorithm can be *consistent* in the sense that for any target function, $f^\star$, the discrepancy between f (which is learned from the training data) and $f^\star$ goes to zero (in some appropriate sense) as the sample size goes to infinity. Since the linear hull of $\{k(\cdot, x) : x \in X\}$ is dense in the RKHS, $\mathcal{H}$ associated with k (Aronszajn, 1950), and assuming that the kernel-based algorithm makes f "converge to an appropriate function" in $\mathcal{H}$ as $n \to \infty$, the above question of approximating $f^\star$ arbitrarily *well* by f in (1) as n goes to infinity is equivalent to the question of whether $\mathcal{H}$ is rich enough to approximate any $f^\star$ arbitrarily *well* (such an RKHS is referred to as a universal RKHS and the corresponding kernel as a universal kernel). Depending on the choice of X, the choice of target function space and the type of approximation, various notions of universality—*c*-universality (Steinwart, 2001), *cc*-universality (Micchelli et al., 2006; Caponnetto et al., 2008), c_0-universality (Carmeli et al., 2010; Sriperumbudur et al., 2010a) and L_p-universality (Steinwart and Christmann, 2008; Carmeli et al., 2010)—have been proposed and characterized in literature.

Recently, a seemingly related (to universality) notion of characteristic kernel has been proposed and characterized (Fukumizu et al., 2004, 2008, 2009; Gretton et al., 2007; Sriperumbudur et al., 2008, 2009, 2010b), which has found applications in testing for homogeneity (Gretton et al., 2007), independence (Gretton et al., 2008), conditional independence (Fukumizu et al., 2008), to find the most predictive subspace in regression (Fukumizu et al., 2004), etc. Formally, given the set of all Borel probability measures defined on the topological space X, a measurable and bounded kernel, k is said to be characteristic if

$$\mathbb{P} \mapsto \int_X k(\cdot, x)\, d\mathbb{P}(x), \tag{2}$$

is injective, that is, $\mathbb{P}$ is embedded to a unique element, $\int_X k(\cdot, x)\, d\mathbb{P}(x)$ in $\mathcal{H}$. The motivation to consider such an embedding is that it provides a powerful and straightforward method of dealing with higher-order statistics of random variables, which has been exploited in the above mentioned applications. Gretton et al. (2007) related characteristic and universal kernels by showing that if k is *c*-universal—see Section 2 for the definition—then it is characteristic. Besides this result, not much is known or understood about the relation between universal and characteristic kernels.

The main contribution of this paper is to clarify the relation between universal and characteristic kernels by presenting a unifying study relating them to RKHS embedding of measures (Suquet, 2009), in addition to clarifying their relation to other common notions of strictly pd, conditionally strictly pd and *integrally strictly pd* kernels, which extends our preliminary study in Sriperumbudur et al. (2010b, Section 3.4). This is done by first reviewing all the existing characterizations for universal and characteristic kernels, which is then used to clarify not only the relation between them but also their relation to other notions of pd kernels (see Section 3). Since the existing characterizations do not explain the complete relationship between all these various notions of pd kernels, we raise open questions in Section 3 about the relationships to be clarified, which are then addressed in Section 4 by deriving new results. In particular, in Section 4, we establish the relation between (a)

c_0-universality and RKHS embedding of finite signed Borel measures, (b) universal and integrally strictly pd kernels, (c) characteristic and conditionally strictly pd kernels and (d) all the above mentioned notions when the pd kernel is *radial* on $\mathbb{R}^d$. A summary of the relation between all these notions of pd kernels is shown in Figure 1, which shows the equivalence between these notions for *radial* kernels on $\mathbb{R}^d$. Supplementary results are collected in appendices. Throughout the paper, we assume X to be a Polish space,[1] the reason for which is discussed in the paragraph following (3).

In the following section, we introduce the notation and collect all definitions that are used throughout the paper.

2. Definitions and Notation

Let X be a topological space. $C(X)$ denotes the space of all continuous real-valued functions on X. $C_b(X)$ is the space of all bounded, continuous real-valued functions on X. For a locally compact Hausdorff space (examples include $\mathbb{R}^d$, infinite discrete sets, topological manifolds, etc.), X, $f \in C(X)$ is said to *vanish at infinity* if for every $\varepsilon > 0$ the set $\{x : |f(x)| \geq \varepsilon\}$ is compact.[2] The class of all continuous f on X which vanish at infinity is denoted as $C_0(X)$. The spaces $C_b(X)$ and $C_0(X)$ are endowed with the uniform norm, $\|\cdot\|_u$ defined as $\|f\|_u := \sup_{x\in X} |f(x)|$ for $f \in C_0(X) \subset C_b(X)$.

Radon measure: A signed Radon measure μ on a Hausdorff space X is a Borel measure on X satisfying

(*i*) $\mu(C) < \infty$ for each compact subset $C \subset X$,

(*ii*) $\mu(B) = \sup\{\mu(C) \,|\, C \subset B, C \text{ compact}\}$ for each B in the Borel σ-algebra of X.

μ is said to be finite if $\|\mu\| := |\mu|(X) < \infty$, where $|\mu|$ is the total-variation of μ. $M_b^+(X)$ denotes the space of all finite Radon measures on X while $M_b(X)$ denotes the space of all finite signed Radon measures on X. The space of all Radon probability measures is denoted as $M_1^+(X) := \{\mu \in M_b^+(X) : \mu(X) = 1\}$. For $\mu \in M_b(X)$, the support of μ is defined as

$$\text{supp}(\mu) = \{x \in X \,|\, \text{for any open set } U \text{ such that } x \in U, |\mu|(U) \neq 0\}. \tag{3}$$

$M_{bc}(X)$ denotes the space of all compactly supported finite signed Radon measures on X. We refer the reader to Berg et al. (1984, Chapter 2) for a general reference on the theory of Radon measures. If X is a Polish space, then by Ulam's theorem, every finite Borel measure is Radon (Dudley, 2002, Theorem 7.1.4). Therefore, for the simplicity of not requiring to distinguish between Borel and Radon measures, throughout the paper, we assume X to be a Polish space.

Positive definite (pd), strictly pd, conditionally strictly pd and integrally strictly pd: A symmetric function $k : X \times X \to \mathbb{R}$ is called positive definite (pd) (*resp.* conditionally pd) if, for all $n \in \mathbb{N}$ (*resp.* $n \geq 2$), $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ (*resp.* with $\sum_{j=1}^n \alpha_j = 0$) and all $x_1, \ldots, x_n \in X$, we have

$$\sum_{l,j=1}^{n} \alpha_l \alpha_j k(x_l, x_j) \geq 0. \tag{4}$$

1. A topological space (X, τ) is called a Polish space if the topology τ has a countable basis and there exists a complete metric defining τ. An example of a Polish space is $\mathbb{R}^d$ endowed with its usual topology.
2. LCH spaces have a rich supply of continuous functions that vanish outside compact sets—see Tietze extension theorem (Folland, 1999, Theorem 4.34).

Furthermore, k is said to be strictly pd (*resp.* conditionally strictly pd) if, for mutually distinct $x_1,\ldots,x_n \in X$, equality in (4) only holds for $\alpha_1 = \cdots = \alpha_n = 0$.

A measurable, symmetric and bounded kernel, k is said to be integrally strictly pd if

$$\iint_X k(x,y)\,d\mu(x)\,d\mu(y) > 0, \forall\, \mu \in M_b(X)\backslash\{0\}.$$

This definition is a generalization of *integrally strictly positive definite functions* on $\mathbb{R}^d$ (Stewart, 1976, Section 6): $\iint_{\mathbb{R}^d} k(x,y)f(x)f(y)\,dx\,dy > 0$ for all $f \in L^2(\mathbb{R}^d)$, which is the strictly positive definiteness of the integral operator given by the kernel.

c-, cc-, c_0- and L_p-universal kernels: A continuous pd kernel k on a compact Hausdorff space X is called *c*-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C(X)$ w.r.t. the uniform norm, that is, for every function $g \in C(X)$ and all $\varepsilon > 0$, there exists an $f \in \mathcal{H}$ such that $\|f-g\|_u \leq \varepsilon$ (Steinwart, 2001).

A continuous pd kernel k on a Hausdorff space X is said to be *cc*-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C(X)$ endowed with the topology of compact convergence, that is, for any compact set $Z \subset X$, for any $g \in C(Z)$ and all $\varepsilon > 0$, there exists an $f \in \mathcal{H}|_Z$ such that $\|f-g\|_u \leq \varepsilon$, where $\mathcal{H}|_Z := \{f|_Z : f \in \mathcal{H}\}$ is the restriction of $\mathcal{H}$ to Z and $f|_Z$ is the restriction of f to Z (Carmeli et al., 2010; Sriperumbudur et al., 2010a).

A pd kernel, k is said to be a c_0-kernel if it is bounded with $k(\cdot,x) \in C_0(X), \forall x \in X$, where X is a locally compact Hausdorff (LCH) space. A c_0-kernel on an LCH space, X is said to be c_0-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C_0(X)$ w.r.t. the uniform norm (Carmeli et al., 2010; Sriperumbudur et al., 2010a).[3]

A measurable and bounded kernel, k defined on a Hausdorff space, X is said to be L_p-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $L^p(X,\mu)$ w.r.t. the *p-norm*, defined as

$$\|f\|_p := \left(\int_X |f(x)|^p\,d\mu(x)\right)^{1/p},$$

for all Borel probability measures, μ, defined on X and some $p \in [1,\infty)$. Here $L^p(X,\mu)$ is the Banach space of p-integrable μ-measurable functions on X (Steinwart and Christmann, 2008).

We would like to stress that in the above definitions of universality, the assumptions on k ensure that the associated RKHS, $\mathcal{H}$ is continuously included in the target space. Steinwart and Christmann (2008, Lemma 4.28) showed that k is bounded and $k(\cdot,x)$ is continuous for all $x \in X$ (X being a topological space) if and only if every $f \in \mathcal{H}$ is bounded and continuous. In addition, the inclusion $\mathrm{id} : \mathcal{H} \to C_b(X)$ is continuous. Similarly, by modifying the proof of Lemma 4.28 in Steinwart and Christmann (2008), it can be easily shown that k is bounded and $k(\cdot,x) \in C_0(X), \forall x \in X$ (X being an LCH space) if and only if every $f \in \mathcal{H}$ is in $C_0(X)$, and the inclusion $\mathrm{id} : \mathcal{H} \to C_0(X)$ can be shown to be continuous (also see Carmeli et al., 2010, Proposition 2.2). Steinwart and Christmann (2008, Theorem 4.26) showed that if k is measurable and bounded on a measurable space X, then

3. Note that *cc*-universality (*resp. c*-universality) deals with X being a non-compact (*resp.* compact) Hausdorff space, whereas c_0-universality requires X to be an LCH space. While X being Hausdorff ensures that it has an abundance of compact subsets (as required in *cc*-universality), the stronger condition of X being an LCH space ensures that it has an abundance of continuous functions that vanish outside compact sets (see footnote 2). In addition, this choice of X being an LCH space ensures the existence of topological dual of $C_0(X)$ through the Riesz representation theorem, which is required in the characterization of c_0-universality. See Proposition 2 in Section 4 for details.

$\mathcal{H}$ consists of p-integrable (w.r.t. any Borel probability measure, μ) functions and the inclusion $\mathrm{id} : \mathcal{H} \to L^p(X,\mu)$ is continuous for some $p \in [1,\infty)$.

Characteristic kernel: A bounded measurable kernel, k is said to be characteristic if $\mu \mapsto \int_X k(\cdot,x)\,d\mu(x)$ is injective, where μ is a Borel probability measure on X.

Translation invariant and Radial kernels on $\mathbb{R}^d$: A pd kernel, $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be translation invariant if $k(x,y) = \psi(x-y)$, where ψ is a pd function. If k is bounded and continuous, then by Bochner's theorem (Wendland, 2005, Theorem 6.6), $\psi \in C_b(\mathbb{R}^d)$ is the Fourier transform of $\Lambda \in M_b^+(\mathbb{R}^d)$, that is,

$$\psi(x) = \int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\,d\Lambda(\omega),\, x \in \mathbb{R}^d. \tag{5}$$

A bounded continuous kernel, k is said to be radial on $\mathbb{R}^d \times \mathbb{R}^d$ if there exists $\nu \in M_b^+([0,\infty))$ such that

$$k(x,y) = \int_{[0,\infty)} e^{-t\|x-y\|_2^2}\,d\nu(t),\, x,y \in \mathbb{R}^d. \tag{6}$$

It is easy to see that a radial kernel is also bounded translation invariant on $\mathbb{R}^d$ (see Appendix A). Examples of radial kernels include the Gaussian kernel, $k(x,y) = e^{-\sigma\|x-y\|_2^2}$, $\sigma > 0$; inverse multiquadrics, $k(x,y) = (c+\|x-y\|_2^2)^{-\beta}$, $\beta > d/2$, etc.

A continuous pd kernel is said to be translation invariant on $\mathbb{T}^d := [0,2\pi)^d$ if $k(x,y) = \psi((x-y)_{\mathrm{mod}\,2\pi})$, where $\psi \in C(\mathbb{T}^d)$ is such that

$$\psi(x) = \sum_{n\in\mathbb{Z}^d} A_\psi(n) e^{\sqrt{-1}x^T n},\, x \in \mathbb{T}^d, \tag{7}$$

with $A_\psi : \mathbb{Z}^d \to \mathbb{R}_+$, $A_\psi(-n) = A_\psi(n)$ and $\sum_{n\in\mathbb{Z}^d} A_\psi(n) < \infty$.

3. Relation Between Various Notions of Positive Definite Kernels Based on Known Characterizations

In this section, we review existing results on the characterization of universal and characteristic kernels, which are then used to clarify not only the relation between them but also their relation to other notions like strictly pd, conditionally strictly pd and integrally strictly pd kernels. In Section 3.1, we discuss various notions of universality, review all their existing characterizations and then summarize the relation between them. In Section 3.2, we discuss and summarize the relation between characteristic and universal kernels based on their existing characterizations. The relation of universal and characteristic kernels to strictly pd, conditionally strictly pd and integrally strictly pd kernels are summarized in Section 3.3. Since the existing characterizations do not explain the complete relationship between all these various notions of pd kernels, we raise questions at the end of each subsection that need to be addressed to obtain a complete understanding of the relationships between all these notions. A summary of the relationships between various notions of pd kernels based on the existing characterizations is shown in Figure 1.

Before proceeding further, we would like to highlight a possible confusion that can raise while comparing these various notions of pd kernels. Suppose we would like to compare c_0-universal vs. characteristic kernels, that is, (a) Is a c_0-universal kernel characteristic? (b) Is the converse true? While (a) is a valid question, answering (b) trivially yields that characteristic kernels are not c_0-

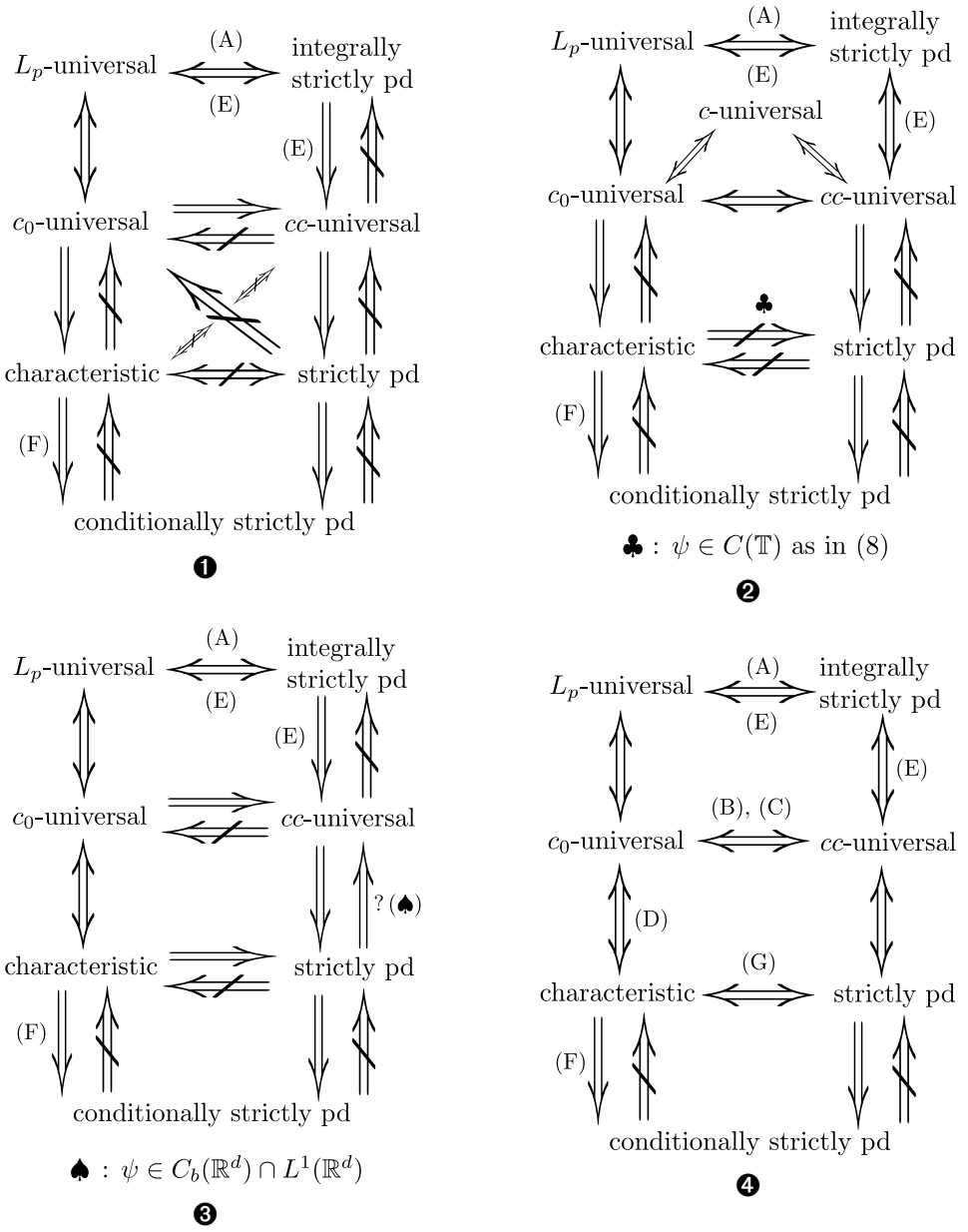

Figure 1: Summary of the relations between various families of c_0-kernels: The implications shown without any reference are based on the review of existing results (see Section 3) while the ones with a reference are based on new results derived in Section 4 that addresses the open questions (A)–(G). The implications which are still open are shown with "?". ❶ X is an LCH space. ❷ The implications shown hold for any compact Hausdorff space, X. When $X = \mathbb{T}$ and k is continuous and translation invariant on $\mathbb{T}$—see (7)—then k being characteristic implies it is strictly pd, which is shown as ♣. ❸ The implications shown hold for bounded continuous translation invariant kernels on $\mathbb{R}^d$—see (5). If $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, then the implication shown as (♠) holds, that is, strictly pd kernels are cc-universal. Otherwise, it is not clear whether the implication holds. ❹ Radial kernels on $\mathbb{R}^d$—see (6).

universal. This is because k need not be a c_0-kernel for it to be characteristic.[4] Therefore, to make a non-trivial comparison between characteristic and c_0-universal kernels, it is important that we assume k to be a c_0-kernel before answering the questions in (a) and (b). In extending this reasoning for the non-trivial comparison of any two notions of pd kernels, it is important to assume that k satisfies the strongest possible condition. Therefore, in order to present a concise summary of the relationships between these various notions, in Figure 1, we assume k to be a c_0-kernel—this is the strongest condition to be satisfied in order to compare all these notions of pd kernels.

3.1 Relation Between Various Notions of Universality

As mentioned before, a universal kernel is such that its corresponding RKHS, $\mathcal{H}$ is rich enough to approximate any target function (belonging to some target space) arbitrarily well. Therefore, depending on the choice of X, the choice of target space and the type of approximation, various notions of universality—*c*, *cc*, c_0 and L^p—have been proposed. In the following, we review the existing characterizations for all these notions of universal kernels and summarize the relation between them.

c-universality: Steinwart (2001) proposed the notion of *c*-universality, wherein X is a compact metric space with $C(X)$ being the target space and $\mathcal{H}$ being dense in $C(X)$ w.r.t. the uniform norm. By applying the Stone-Weierstraß theorem (Folland, 1999, Theorem 4.45), Steinwart (2001, Theorem 9) provided sufficient conditions for a kernel to be *c*-universal—a continuous kernel, k on a compact metric space, X is *c*-universal if the following hold: (a) $k(x,x) > 0, \forall x \in X$, (b) there exists an injective feature map $\Phi : X \to \ell_2$ of k with $\Phi(x) = \{\Phi_n(x)\}_{n\in\mathbb{N}}$ and (c) $\text{span}\{\Phi_n : n \in \mathbb{N}\}$ is an algebra—using which the Gaussian kernel is shown to be *c*-universal on every compact subset of $\mathbb{R}^d$. Micchelli et al. (2006, Proposition 1) related *c*-universality to the injective RKHS embedding of finite signed Borel measures by showing that k is *c*-universal if and only if

$$\mu \mapsto \int_X k(\cdot,x)\, d\mu(x),\ \mu \in M_b(X), \tag{8}$$

is injective.

cc-universality: One limitation in the notion of universality considered by Steinwart (2001) is that X is assumed to be compact, which excludes many interesting spaces, such as $\mathbb{R}^d$ and infinite discrete sets. To overcome this limitation, Carmeli et al. (2010, Definition 4.1, Theorem 4.3) and Sriperumbudur et al. (2010a) introduced the notion of *cc*-universality which can handle non-compact Hausdorff spaces, X. Carmeli et al. (2010, Proposition 2.3, Theorems 4.3 and 4.4) showed that a bounded continuous pd kernel, k is *cc*-universal if and only if the following embedding is injective for all $\mu \in M_{bc}(X)$ and some $p \in [1,\infty)$:

$$f \mapsto \int_X k(\cdot,x)\, f(x)\, d\mu(x),\ f \in L^p(X,\mu). \tag{9}$$

In addition, Carmeli et al. (2010, Remark 4.1) showed that k being *cc*-universal is equivalent to it being universal in the sense of Micchelli et al. (2006) and Caponnetto et al. (2008): for any compact $Z \subset X$, the set $K(Z) := \overline{\text{span}}\{k(\cdot,y) : y \in Z\}$ is dense in $C(Z)$ in the uniform norm, which is shown by

4. Let k_1 be a characteristic kernel on $\mathbb{R}$. Define $k_2(x,y) = 1$ if $x = y \in \mathbb{R}$ and $k_2(x,y) = 0$ if $x \neq y \in \mathbb{R}$. Clearly k_2 is not continuous and therefore $k_1 + k_2$ is not a c_0-kernel, even if k_1 is a c_0-kernel. However, it is easy to verify that $k_1 + k_2$ is characteristic.

Micchelli et al. (2006, Proposition 1) to be equivalent to the following embedding being injective:

$$\mu \mapsto \int_Z k(\cdot,x)\, d\mu(x),\ \mu \in M_b(Z). \tag{10}$$

Since (10) holds for any compact $Z \subset X$, the universality in the sense of Micchelli et al. and Caponnetto et al. is equivalent to the following embedding being injective:

$$\mu \mapsto \int_X k(\cdot,x)\, d\mu(x),\ \mu \in M_{bc}(X). \tag{11}$$

Therefore, k being *cc*-universal is equivalent to the injectivity of (11)—in Section 4, we present a more direct proof of this result (see Remark 3). It is clear from the definitions of *c*- and *cc*-universality that these notions are equivalent when X is compact, which also follows from their characterizations in (8) and (11).

As special cases, Micchelli et al. (2006, Propositions 14, Theorem 17) showed that a translation invariant kernel on $\mathbb{R}^d$ is *cc*-universal if $\text{supp}(\Lambda)$ is a uniqueness subset[5] of $\mathbb{C}^d$, while a radial kernel on $\mathbb{R}^d$ is *cc*-universal if and only if $\text{supp}(\nu) \neq \{0\}$—see (5) and (6) for the definitions of Λ and ν. Using these characterizations, many popular kernels on $\mathbb{R}^d$ are shown to be *cc*-universal (Micchelli et al., 2006, Section 4): Gaussian, Laplacian, B_{2l+1}-spline, sinc kernel, etc.

c_0-universality: Although *cc*-universality solves the limitation of *c*-universality by handling non-compact X, the topology of compact convergence considered in *cc*-universality is weaker than the topology of uniform convergence, that is, a sequence of functions, $\{f_n\} \subset C(X)$ converging to $f \in C(X)$ in the topology of uniform convergence ensures that they converge in the topology of compact convergence but not vice-versa. So, the natural question to ask is whether we can characterize $\mathcal{H}$ that are rich enough to approximate any $f^\star$ on non-compact X in a stronger sense, that is, uniformly, by some $g \in \mathcal{H}$. Carmeli et al. (2010, Definition 2.2, Theorem 4.1) and Sriperumbudur et al. (2010a) answered this through the notion of c_0-universality, wherein X is an LCH space with $C_0(X)$ being the target space and $\mathcal{H}$ being dense in $C_0(X)$ w.r.t. the uniform norm (note that a notion of universality that is stronger than c_0-universality can be defined by choosing X to be a Hausdorff space, $C_b(X)$ to be the target space and $\mathcal{H}$ being dense in $C_b(X)$ w.r.t. the uniform norm. However, this notion of universality does not enjoy a nice characterization as c_0-universality—see (12) and (13) for the characterization of c_0-universality—and therefore, we did not include it in our study of relationships between various notions of pd kernels. See Appendix C for details).

Carmeli et al. (2010, Theorem 4.1) showed that a c_0-kernel k is c_0-universal if and only if it is L_p-universal, which by Proposition 2.3 and Theorem 4.2 of Carmeli et al. (2010) is equivalent to the injectivity of the following embedding for all $\mu \in M_b(X)$ and some $p \in [1,\infty)$:

$$f \mapsto \int_X k(\cdot,x)\, f(x)\, d\mu(x),\ f \in L^p(X,\mu). \tag{12}$$

We provide an alternate characterization for c_0-universality in Section 4 (see Proposition 2) that k is c_0-universal if and only if the following embedding is injective:

$$\mu \mapsto \int_X k(\cdot,x)\, d\mu(x),\ \mu \in M_b(X). \tag{13}$$

5. A subset S of $\mathbb{C}^d$ is a uniqueness set if an entire function on $\mathbb{C}^d$ vanishes on S then it is everywhere zero on $\mathbb{C}^d$. Non-empty interior is sufficient for a set to be a uniqueness set.

As a special case, Carmeli et al. (2010, Proposition 5.6) showed that a translation invariant k on $\mathbb{R}^d$ is c_0-universal if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$. Examples of c_0-universal kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, B_{2l+1}-spline, inverse multiquadrics, Matérn class, etc.

Summary: The following statements summarize the relation between various notions of universality, which are depicted in Figure 1.

- c- and cc-universality are related to the injective RKHS embedding of finite signed Borel measures, as shown in (8) and (11).
- For c_0-kernels defined on an LCH space X, c_0-universality implies cc-universality, which follows from (9) and (12). The converse is however not true as a bounded continuous translation invariant c_0-kernel on $\mathbb{R}^d$ is c_0-universal if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ while $(\text{supp}(\Lambda))^\circ \neq \emptyset$ is sufficient for cc-universality, where A° represents the interior of A.
- When X is compact, then c-, cc- and c_0-universality are equivalent.
- For an LCH space X, a c_0-kernel is c_0-universal if and only if it is L_p-universal.
- If k is a radial kernel on $\mathbb{R}^d$, then k is cc-universal if and only if $\text{supp}(\nu) \neq \{0\}$.

Open questions: The following relationships need to be clarified, which we do in Section 4.

(A) As mentioned in the summary, c- and cc-universality are related to the injective RKHS embedding of finite signed Borel measures. However, the relation between c_0-universality and the injective RKHS embedding of finite signed Borel measures as shown in (13) is not clear, which we clarify in Section 4.1.

(B) For c_0-kernels defined on an LCH space X (that is not compact), it is clear from the summary that c_0-universality implies cc-universality. Is there a case for which cc-universality implies c_0-universality? We address this in Section 4.3.

(C) While cc-universality is characterized for radial kernels on $\mathbb{R}^d$, the characterization of c_0-universality for radial kernels is not known. In Section 4.3, we provide a characterization of c_0-universality for radial kernels on $\mathbb{R}^d$ and then establish the relation between c_0-universality and cc-universality for such kernels.

3.2 Relation Between Characteristic and Universal Kernels

In this section, we comprehensively clarify the relation between various notions of universality and characteristic kernels, based on already existing characterizations for characteristic kernels and the results summarized in Section 3.1 for universal kernels.

c-universal kernels vs. Characteristic kernels: Gretton et al. (2007) related universal and characteristic kernels by showing that if k is c-universal, then it is characteristic. In our preliminary study in Sriperumbudur et al. (2010b, Section 3.4), we showed that the converse is not true: as an example, a translation invariant kernel, k on $\mathbb{T}^d \times \mathbb{T}^d$ is characteristic if and only if $A_\psi(0) \geq 0$, $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}_+^d$ while it is universal if and only if $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$.

cc-universal kernels vs. Characteristic kernels: cc-universal kernels on a non-compact Hausdorff space need not be characteristic: for example, a bounded continuous translation invariant

kernel on $\mathbb{R}^d$ is *cc*-universal if $(\text{supp}(\Lambda))^\circ \neq \emptyset$ (see the summary of Section 3.1) while it is characteristic if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ (Sriperumbudur et al., 2008, Theorem 7). Although, this example shows that a bounded continuous translation invariant kernel on $\mathbb{R}^d$ is *cc*-universal if it is characteristic, it is not clear whether such a relation holds on a general non-compact Hausdorff space (not necessarily $\mathbb{R}^d$). The following example shows that continuous kernels that are characteristic on non-compact Hausdorff space, X also need not be *cc*-universal.

Example 1 *Let $X = \mathbb{N}$. Define $k(x,y) = \delta_{xy}, x,y \in X\backslash\{1\}$, $k(x,1) = 0$ for any $x \in X$, where δ represents the Kronecker delta. Suppose $\mu = \delta_1 \in M_{bc}(X)\backslash\{0\}$, where δ_j represents the Dirac measure at j. Then $\|\int_X k(\cdot,x)\,d\mu(x)\|^2_{\mathcal{H}} = \|k(\cdot,1)\|^2_{\mathcal{H}} = k(1,1) = 0$, which means there exists $\mu \in M_{bc}(X)\backslash\{0\}$ such that $\int_X k(\cdot,x)\,d\mu(x) = 0$, that is, (11) is not injective and therefore k is not cc-universal. However, k is characteristic as we show below.*

Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on X such that $\mathbb{P} = \sum_{j\in\mathbb{N}} p_j\delta_j$, $\mathbb{Q} = \sum_{j\in\mathbb{N}} q_j\delta_j$ with $p_j \geq 0, q_j \geq 0$ for all $j \in \mathbb{N}$ and $\sum_{j\in\mathbb{N}} p_j = \sum_{j\in\mathbb{N}} q_j = 1$. Consider

$$\begin{aligned} B &:= \left\| \int_X k(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\right\|^2_{\mathcal{H}} = \left\| \sum_{j\in\mathbb{N}} (p_j - q_j)k(\cdot,j)\right\|^2_{\mathcal{H}} = \sum_{l,j\in\mathbb{N}} (p_l - q_l)(p_j - q_j)k(l,j) \\ &= (p_1-q_1)^2k(1,1) + 2(p_1-q_1)\sum_{j\in\mathbb{N}\backslash\{1\}} (p_j-q_j)k(j,1) + \sum_{l,j\in\mathbb{N}\backslash\{1\}} (p_j-q_j)(p_l-q_l)k(j,l) \\ &= \sum_{j\in\mathbb{N}\backslash\{1\}} (p_j-q_j)^2. \end{aligned}$$

Suppose $B = 0$, which means $p_j = q_j, \forall\, j \in \mathbb{N}\backslash\{1\}$. Since $\sum_{j\in\mathbb{N}} p_j = \sum_{j\in\mathbb{N}} q_j = 1$, we have $p_1 = q_1$ and so $\mathbb{P} = \mathbb{Q}$, that is, (2) is injective and therefore k is characteristic.

c_0-universal kernels vs. Characteristic kernels: Fukumizu et al. (2008, 2009) have shown that a measurable and bounded kernel, k is characteristic if and only if $\mathcal{H} + \mathbb{R}$ (the direct sum of $\mathcal{H}$ and $\mathbb{R}$ is defined as $\mathcal{H} + \mathbb{R} := \{f + c : f \in \mathcal{H}, c \in \mathbb{R}\}$) is dense in $L^p(X,\mathbb{P})$ for all $\mathbb{P} \in M_1^+(X)$ and for some $p \in [1,\infty)$. Using this, it is easy to see that if $\mathcal{H}$ is dense in $L^p(X,\mathbb{P})$ for all $\mathbb{P} \in M_1^+(X)$ and for some $p \in [1,\infty)$, then k is characteristic. Based on the results summarized in Section 3.1, it is clear that for an LCH space, X, if k is c_0-universal, which means k is L_p-universal, then $\mathcal{H}$ is dense in $L^p(X,\mathbb{P})$ for all $\mathbb{P} \in M_1^+(X)$ and for some $p \in [1,\infty)$ and therefore is characteristic. In Section 4, we provide an alternate proof for this relation between c_0-universal and characteristic kernels by answering (A). Clearly, the converse is not true, that is, a c_0-kernel that is characteristic need not be c_0-universal (see Proposition 4 and footnote 8). However, for bounded continuous translation invariant kernels on $\mathbb{R}^d$, the converse is true, that is, a translation invariant c_0-kernel that is characteristic[6] is also c_0-universal. This is because of the fact that a translation invariant kernel on $\mathbb{R}^d$ is characteristic if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ (Sriperumbudur et al., 2008, Theorem 7), which is also the same characterization summarized in Section 3.1 for c_0-universal kernels.

Summary: The following statements summarize the relation between universal and characteristic kernels, which are depicted in Figure 1.

6. Let $k(x,y) = \psi(x-y)$ be a bounded continuous translation invariant kernel on $\mathbb{R}^d$, which by Bochner's theorem is of the form in (5). Suppose $\psi \in L^1(\mathbb{R}^d)$. Then by the Fourier inversion theorem (Dudley, 2002, Theorem 9.5.4), Λ has a density, $\hat{\psi}$ w.r.t. the Lebesgue measure such that $\hat{\psi} \in L^1(\mathbb{R}^d)$. Therefore, since ψ is the Fourier transform of $\hat{\psi}$, by the Riemann-Lebesgue lemma (Rudin, 1991, Theorem 7.5), $\psi \in C_0(\mathbb{R}^d)$, that is, k is a c_0-kernel. Most of the well-known characteristic kernels satisfy the condition of $\psi \in L^1(\mathbb{R}^d)$ and therefore are c_0-kernels. This means, for all practical purposes, we can assume bounded continuous translation invariant kernels to be c_0-kernels.

- For c_0-kernels defined on an LCH space, X, L_p-universal $\Leftrightarrow$ c_0-universal $\Rightarrow$ characteristic. But in general, c_0-kernels that are characteristic need not be c_0-universal. However, for translation invariant kernels on $\mathbb{R}^d$, c_0-universal $\Leftrightarrow$ characteristic.
- When X is compact, c-universal $\Rightarrow$ characteristic but not vice-versa.
- For translation invariant kernels on $\mathbb{R}^d$, characteristic $\Rightarrow$ cc-universal but not vice-versa. However, on general non-compact Hausdorff spaces, continuous kernels that are characteristic need not be cc-universal.

Open questions: The following relationship need to be clarified, which we do in Section 4.

(D) While the relation between universal and characteristic kernels that are translation invariant on $\mathbb{R}^d$ is clear (see the summary above), the characterization of characteristic and c_0-universal kernels that are radial on $\mathbb{R}^d$ is not known and therefore the relation between characteristic and universal kernels that are radial on $\mathbb{R}^d$ is not clear. We address this in Section 4.3.

3.3 Relation of Universal and Characteristic Kernels to Strictly PD, Integrally Strictly PD and Conditionally Strictly PD Kernels

In this section, we relate characteristic kernels and various notions of universal kernels to strictly pd, integrally strictly pd and conditionally strictly pd kernels. Before that, we summarize the relation between strictly pd, integrally strictly pd and conditionally strictly pd kernels. In Sriperumbudur et al. (2010b, Section 3.4), we showed that integrally strictly pd kernels are strictly pd. The converse is not true, which follows from Steinwart and Christmann (2008, Proposition 4.60, Theorem 4.62). However, if X is a finite set, then k being strictly pd also implies it is integrally strictly pd. From the definitions of strictly pd and conditionally strictly pd kernels, it is clear that a strictly pd kernel is conditionally strictly pd but not vice-versa.

Universal kernels vs. Strictly pd kernels: Carmeli et al. (2010, Corollary 4.3) showed that cc-universal kernels are strictly pd, which means c_0-universal kernels are also strictly pd (as c_0-universal $\Rightarrow$ cc-universal from Section 3.1). This means, when X is compact Hausdorff, c-universal kernels are strictly pd, which matches with the result in Steinwart and Christmann (2008, Definition 4.53, Proposition 4.54, Example 4.11).

Conversely, a strictly pd c_0-kernel on an LCH space need not be c_0-universal. This follows from Theorem 4.62 in Steinwart and Christmann (2008) which shows that there exists a bounded strictly pd kernel, k on $X := \mathbb{N} \cup \{0\}$ with $k(\cdot, x) \in C_0(X), \forall x \in X$ such that k is not L_p-universal (which from the summary of Section 3.1 means k is not c_0-universal). Similarly, when X is compact, the converse is not true, that is, continuous strictly pd kernels need not be c-universal which follows from the results due to Dahmen and Micchelli (1987) and Pinkus (2004) for Taylor kernels (Steinwart and Christmann, 2008, Lemma 4.8, Corollary 4.57)—refer to Steinwart and Christmann (2008, Section 4.7, p. 161) for more details.[7] Therefore, it is evident that a continuous strictly pd kernel is in general not cc-universal on an Hausdorff space. However, for translation invariant kernels that are continuous, bounded and integrable on $\mathbb{R}^d$, that is, $k(x,y) = \psi(x-y), x, y \in \mathbb{R}^d$, where $\psi \in$

7. Another example of continuous strictly pd kernels that are not c-universal is as follows. Using the technique in the proof of Theorem 14 of Sriperumbudur et al. (2010b), it can be shown that a continuous translation invariant kernel on $\mathbb{T} \times \mathbb{T}$ is c-universal if and only if $A_\psi(n) > 0, \forall n \in \mathbb{Z}$. Therefore, by Theorem 8 (see Appendix B), a strictly pd kernel on $\mathbb{T}$ need not be c-universal.

$C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, strictly pd implies *cc*-universality. This follows from Theorem 6.11 and Corollary 6.12 of Wendland (2005) that if $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ is strictly pd, then $(\text{supp}(\Lambda))^\circ \neq \emptyset$, which from the summary of Section 3.1 means k is *cc*-universal. Similarly, when the kernel is radial on $\mathbb{R}^d$, then strictly pd kernels are *cc*-universal. This follows from Theorem 7.14 of Wendland (2005), which shows that a radial kernel on $\mathbb{R}^d$ is strictly pd if and only if $\text{supp}(\nu) \neq \{0\}$, and therefore *cc*-universal (from the summary of Section 3.1). On the other hand, when X is finite, all these notions of universal and strictly pd kernels are equivalent, which follows from the result due to Carmeli et al. (2010, Corollary 4.3) that *cc*-universal and strictly pd kernels are the same when X is finite.

Characteristic kernels vs. Strictly pd kernels: Since characteristic kernels that are c_0- and translation invariant on $\mathbb{R}^d$ are equivalent to c_0-universal kernels (see the summary of Section 3.2), it is clear that they are strictly pd. However, the converse is not true: for example, the sinc-squared kernel, $k(x,y) = \frac{\sin^2(\sigma(x-y))}{(x-y)^2}$ on $\mathbb{R}$, which has $\text{supp}(\Lambda) = [-\sigma,\sigma] \subsetneq \mathbb{R}$ is strictly pd (Wendland, 2005, Theorem 6.11), while it is not characteristic. Based on Example 1, it can be shown that in general, characteristic kernels on a non-compact space (not necessarily $\mathbb{R}^d$) need not be strictly pd: in Example 1, k is characteristic but is not strictly pd because for $(a_1,\ldots,a_n) = (1,0,\ldots,0)$ and $(x_1,\ldots,x_n) = (1,\ldots,n)$, we have $\sum_{l,j=1}^n a_l a_j k(x_l,x_j) = a_1^2 k(1,1) + 2a_1 \sum_{j=2}^n a_j k(j,1) + \sum_{j=2}^n a_j^2 = 0$. Note that Example 1 holds even if X is a compact subset of $\mathbb{N}$. Therefore, when X is compact Hausdorff, a characteristic kernel need not be strictly pd. However, for translation invariant kernels on $\mathbb{T}$, a characteristic kernel is also strictly pd, while the converse is not true: Fukumizu et al. (2009, Theorem 8) and Sriperumbudur et al. (2010b, Theorem 14) have shown that k on $\mathbb{T}\times\mathbb{T}$ is characteristic if and only if $A_\psi(0) \geq 0, A_\psi(n) > 0, \forall n \in \mathbb{Z}\backslash\{0\}$, which by Theorem 8 (see Appendix B) is strictly pd, while the converse is clearly not true.

Characteristic kernels vs. Integrally strictly pd kernels: In Sriperumbudur et al. (2009, Theorem 4) and Sriperumbudur et al. (2010b, Theorem 7), we have shown that integrally strictly pd kernels are characteristic, while the converse in general is not true.[8] When k is bounded continuous and translation invariant on $\mathbb{R}^d$, however the converse holds, which is due to the fact that if k is characteristic, then $\text{supp}(\Lambda) = \mathbb{R}^d$ (Sriperumbudur et al., 2008, Theorem 7), which ensures that k is integrally strictly pd.

Summary: The following statements summarize the relation of universal and characteristic kernels to strictly pd, integrally strictly pd and conditionally strictly pd kernels, which are depicted in Figure 1.

- *c*-, *cc*- and c_0-universal kernels are strictly pd and are therefore conditionally strictly pd, while the converse in general is not true. When X is finite, then *c*-, *cc*- and c_0-universal kernels are equivalent to strictly pd kernels.

- Bounded, continuous, integrable, strictly pd translation invariant kernels on $\mathbb{R}^d$ are *cc*-universal. Radial kernels on $\mathbb{R}^d$ are strictly pd if and only if they are *cc*-universal.

- For a general non-compact Hausdorff space, characteristic kernels need not be strictly pd and vice-versa. However, bounded continuous translation invariant kernels on $\mathbb{R}^d$ or $\mathbb{T}$ that are characteristic are strictly pd but the converse is not true.

8. By Example 1, it is clear that for $\mu = \delta_1 \in M_b(X)\backslash\{0\}$, $\int\int_X k(x,y)\,d\mu(x)\,d\mu(y) = k(1,1) = 0$, where δ_1 represents the Dirac measure at 1. Therefore k is not integrally strictly pd but is characteristic.

- Integrally strictly pd kernels are characteristic. Though the converse is not true in general, it holds if the kernel is bounded, continuous and translation invariant on $\mathbb{R}^d$.

Open questions: The following questions need to be clarified, which is done in Section 4.

(E) While the relation of universal kernels to strictly pd and conditionally strictly pd kernels is clear from the above summary, the relation between universal and integrally strictly pd kernels is not known, which we establish in Section 4.2.

(F) When X is a finite set, it is easy to see that characteristic and conditionally strictly pd kernels are equivalent (see Section 4.4). However, their relationship is not clear for a general measurable space, which we clarify in Section 4.4.

(G) As summarized above, radial kernels on $\mathbb{R}^d$ are strictly pd if and only if they are *cc*-universal. However, the relation between all the other notions of pd kernels—c_0-universal, characteristic, strictly pd and integrally strictly pd—is not known, which is addressed in Section 4.3.

4. Relation Between Various Notions of Positive Definite Kernels: New Results

In this section, we address the open questions, (A)–(G) mentioned in Section 3 to understand the complete relationship between various notions of positive definite kernels.

4.1 c_0-universality and RKHS Embedding of Measures

As mentioned in Section 3.1, Micchelli et al. (2006) have established the relation of *c*-universality and *cc*-universality to injective RKHS embedding of finite signed Borel measures—shown in (8) and (11)—through a simple application of the Hahn-Banach theorem (see Theorem 1). The following result (also see Suquet, 2009, Remark 1.1) in Proposition 2 provides a measure embedding characterization—shown in (13)—for c_0-universality, which is also obtained as a simple application of the Hahn-Banach theorem, and therefore addresses the open question in (A). Before we state Proposition 2, we present the Hahn-Banach theorem, which we quote from Rudin (1991, Theorem 3.5 and the remark following Theorem 3.5).

Theorem 1 (Hahn-Banach) *Suppose A is a subspace of a locally convex topological vector space Y. Then A is dense in Y if and only if $A^\perp = \{0\}$, where*

$$A^\perp := \{T \in Y' : \forall x \in A,\ T(x) = 0\}.$$

The following result, which presents a necessary and sufficient condition for k to be c_0-universal hinges on the above theorem, where we choose A to be the RKHS, $\mathcal{H}$ and Y to be $C_0(X)$ for which Y' is known through the Riesz representation theorem (Folland, 1999, Theorem 7.17).

Proposition 2 (c_0-universality and RKHS embedding of measures) *Suppose X is an LCH space with the kernel, k being bounded and $k(\cdot, x) \in C_0(X), \forall x \in X$. Then k is c_0-universal if and only if the embedding,*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x),\ \mu \in M_b(X), \tag{14}$$

is injective.

Proof By definition, k is c_0-universal if $\mathcal{H}$ is dense in $C_0(X)$. We now invoke Theorem 1 to characterize the denseness of $\mathcal{H}$ in $C_0(X)$, which means we need to consider the dual $C_0'(X) := (C_0(X))'$ of $C_0(X)$. By the Riesz representation theorem (Folland, 1999, Theorem 7.17), $C_0'(X) = M_b(X)$ in the sense that there is a bijective linear isometry $\mu \mapsto T_\mu$ from $M_b(X)$ onto $C_0'(X)$, given by the natural mapping, $T_\mu(f) = \int_X f\, d\mu,\ f \in C_0(X)$. Therefore, by Theorem 1, $\mathcal{H}$ is dense in $C_0(X)$ if and only if $\mathcal{H}^\perp := \{\mu \in M_b(X) : \forall\, f \in \mathcal{H},\ \int_X f\, d\mu = 0\} = \{0\}$. From Lemma 7 (see Appendix B), we have $\mathcal{H}^\perp = \{\mu \in M_b(X) : \int_X k(\cdot,x)\, d\mu(x) = 0\}$ and therefore the result follows from Theorem 1. ■

Remark 3 (a) *When X is compact, $C_0(X)$ coincides with $C(X)$, and therefore the result in (14) matches with the one in (8), derived by Micchelli et al. (2006).*

(b) *The characterization of cc-universality, shown in (11) can also be directly obtained as a simple application of Theorem 1, wherein the proof is similar to that of Proposition 2 except that we need to consider the dual of $C(X)$ endowed with the topology of compact convergence (a locally convex topological vector space) to characterize the denseness of $\mathcal{H}$ in $C(X)$. It is known (Hewitt, 1950) that $C'(X) = M_{bc}(X)$ in the sense that there is a bijective linear isometry $\mu \mapsto T_\mu$ from $M_{bc}(X)$ onto $C'(X)$, given by the natural mapping, $T_\mu(f) = \int_X f\, d\mu,\ f \in C(X)$. The rest of the proof is verbatim with $M_b(X)$ replaced by $M_{bc}(X)$.*

(c) *Comparing (14) and (2), it is clear that c_0-universal kernels are characteristic while the converse is not true, which matches with the result in Section 3.2.*

4.2 Relation Between Universal Kernels and Integrally Strictly PD Kernels

In this section, we address the open question (E) through the following result which shows that c_0-kernels are integrally strictly pd if and only if they are c_0-universal.

Proposition 4 (c_0-universal and integrally strictly pd kernels) *Suppose the assumptions in Proposition 2 hold. Then, a c_0-kernel, k is c_0-universal if and only if it is integrally strictly pd, that is,*

$$\int\int_X k(x,y)\, d\mu(x)\, d\mu(y) > 0, \forall\, \mu \in M_b(X)\backslash\{0\}. \tag{15}$$

Proof ($\Leftarrow$) Suppose k is not c_0-universal. By Proposition 2, there exists $\mu \in M_b(X)\backslash\{0\}$ such that $\int_X k(\cdot,x)\, d\mu(x) = 0$, which implies $\| \int_X k(\cdot,x)\, d\mu(x)\|_{\mathcal{H}} = 0$. This means

$$0 = \left\langle \int_X k(\cdot,x)\, d\mu(x), \int_X k(\cdot,x)\, d\mu(x) \right\rangle_{\mathcal{H}} \overset{(e)}{=} \int\int_X k(x,y)\, d\mu(x)\, d\mu(y),$$

that is, k is not integrally strictly pd, where (e) follows from Lemma 7 (see Appendix B). Therefore, if (15) holds, then k is c_0-universal.

($\Rightarrow$) Suppose there exists $\mu \in M_b(X)\backslash\{0\}$ such that $\int\int_X k(x,y)\, d\mu(x)\, d\mu(y) = 0$, that is,

$$\left\| \int_X k(\cdot,x)\, d\mu(x) \right\|_{\mathcal{H}} = 0 \Rightarrow \int_X k(\cdot,x)\, d\mu(x) = 0.$$

Therefore, the embedding in (14) is not injective, which by Proposition 2 implies that k is not c_0-universal. Therefore, if k is c_0-universal, then k satisfies (15). ■

4.3 Radial Kernels on $\mathbb{R}^d$

In this section, we address the open questions (B), (C), (D) and (G) by showing that all the notions of universality and characteristic kernels are equivalent to strictly pd kernels.

Proposition 5 (All notions are equivalent for radial kernels on $\mathbb{R}^d$) *Suppose k is radial on $\mathbb{R}^d$. Then the following conditions are equivalent.*

(a) $\text{supp}(\nu) \neq \{0\}$.

(b) k is integrally strictly pd.

(c) k is c_0-universal.

(d) k is cc-universal.

(e) k is strictly pd.

(f) k is characteristic.

Proof Note that $(b) \Leftrightarrow (c)$ follows from Proposition 4, $(c) \Rightarrow (d)$ from (11) and (13) and $(d) \Leftrightarrow (e)$ from Micchelli et al. (2006, Proposition 14) and Wendland (2005, Theorem 7.14). Theorem 7.14 in Wendland (2005) also ensures that $(e) \Rightarrow (a)$. Now, we show $(a) \Rightarrow (b)$. To do this, we first derive an intermediate result. Suppose $\hat{\mu}$ is the Fourier transform of μ defined as $\hat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\omega^T x} d\mu(x)$, then for any ψ defined as in (5), we have

$$\begin{aligned}
\int\int_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) &= \int\int\int_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega} d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&= \int\int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega} d\mu(x) \int_{\mathbb{R}^d} e^{\sqrt{-1}y^T\omega} d\mu(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega). \qquad (16)
\end{aligned}$$

Consider $\int\int_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y)$ with k as in (6), given by

$$\begin{aligned}
B := \int\int_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y) &= \int\int_{\mathbb{R}^d} \int_0^\infty e^{-t\|x-y\|_2^2} d\nu(t)\, d\mu(x)\, d\mu(y) \\
&\overset{(\star)}{=} \int_0^\infty \left[\int\int_{\mathbb{R}^d} e^{-t\|x-y\|_2^2} d\mu(x)\, d\mu(y) \right] d\nu(t) \\
&\overset{(\clubsuit)}{=} \int_0^\infty \frac{1}{(4\pi t)^{d/2}} \left[\int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2 e^{-\frac{\|\omega\|_2^2}{4t}} d\omega \right] d\nu(t) \\
&\overset{(\spadesuit)}{=} \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2 \left[\int_0^\infty \frac{1}{(4\pi t)^{d/2}} e^{-\frac{\|\omega\|_2^2}{4t}} d\nu(t) \right] d\omega, \qquad (17)
\end{aligned}$$

where Fubini's theorem is invoked in $(\star)$ and $(\spadesuit)$, while we used (16) in $(\clubsuit)$, where we set $\psi(x) = e^{-t\|x\|_2^2}$ with $d\Lambda(\omega) = (4\pi t)^{-d/2} e^{-\|\omega\|_2^2/4t}\, d\omega$. Since $\text{supp}(\nu) \neq \{0\}$, the inner integral in (17) is positive for every $\omega \in \mathbb{R}^d$ and so $B > 0$, which means k is integrally strictly pd.

We now prove that $(c) \Leftrightarrow (f)$. $(c) \Rightarrow (f)$ follows from Section 3.2. To prove the converse, we need to prove that if k is not c_0-universal, then it is not characteristic. If k is not c_0-universal, then we have $\text{supp}(\nu) = \{0\}$, which means the kernel is a constant function on $\mathbb{R}^d \times \mathbb{R}^d$ and therefore not characteristic. ■

4.4 Relation Between Characteristic and Conditionally Strictly PD Kernels

In this section we address the open question (F) which is about the relation of characteristic kernels to conditionally strictly pd kernels.

As shown in Section 3.3, although the relation between universal and conditionally strictly pd kernels straightforwardly follows from universal kernels being strictly pd, which in turn are conditionally strictly pd, such an implication is not possible in the case of characteristic kernels as they are not in general strictly pd (see Example 1). However, the following result establishes the relation between characteristic and conditionally strictly pd kernels.

Proposition 6 *If k is characteristic, then it is conditionally strictly pd.*

Proof Suppose k is not conditionally strictly pd. This means for some $n \geq 2$ and for mutually distinct $x_1, \ldots, x_n \in X$, there exists $\{\alpha_j\}_{j=1}^n \neq 0$ with $\sum_{j=1}^n \alpha_j = 0$ such that $\sum_{l,j=1}^n \alpha_l \alpha_j k(x_l, x_j) = 0$. Define $I := \{j : \alpha_j > 0\}$, $\mathbb{P} := \beta^{-1} \sum_{j \in I} \alpha_j \delta_{x_j}$ and $\mathbb{Q} := -\beta^{-1} \sum_{j \notin I} \alpha_j \delta_{x_j}$, where $\beta := \sum_{j \in I} \alpha_j$. It is easy to see that $\mathbb{P}$ and $\mathbb{Q}$ are distinct Borel probability measures on X. Then, we have

$$\left\| \int_X k(\cdot, x)\, d(\mathbb{P} - \mathbb{Q})(x) \right\|_{\mathcal{H}}^2 = \beta^{-2} \left\| \sum_{j=1}^n \alpha_j k(\cdot, x_j) \right\|_{\mathcal{H}} = \beta^{-2} \sum_{l,j=1}^n \alpha_l \alpha_j k(x_l, x_j) = 0.$$

So, there exist $\mathbb{P} \neq \mathbb{Q}$ such that $\int_X k(\cdot, x)\, d(\mathbb{P} - \mathbb{Q})(x) = 0$, that is, k is not characteristic. ■

The converse to Proposition 6 in general is however not true: we showed in Section 3.3 that strictly pd kernels are conditionally strictly pd but need not be characteristic and so conditionally strictly pd kernels need not have to be characteristic. In the following, we present a concrete example to show the same—a similar example is used to prove Theorem 4.62 in Steinwart and Christmann (2008), which shows that c_0-kernels that are strictly pd need not be c_0-universal.

Example 2 *Let $X = \mathbb{N} \cup \{0\}$. Define $k(0,0) = \sum_{n \in \mathbb{N}} b_n^2$, $k(m,n) = \delta_{mn}$ and $k(n,0) = b_n$ for $m, n \geq 1$, where $\{b_n\}_{n \geq 1} \subset (0,1)$ and $\sum_{n \in \mathbb{N}} b_n = 1$. Let $n \geq 2$ and $\alpha := (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1}$ be a vector with $\alpha \neq 0$ such that $\sum_{j=0}^n \alpha_j = 0$. Consider*

$$\begin{aligned} B := \sum_{l,j=0}^n \alpha_l \alpha_j k(l,j) &= \alpha_0^2 k(0,0) + 2 \sum_{j=1}^n \alpha_j \alpha_0 k(j,0) + \sum_{l,j=1}^n \alpha_l \alpha_j k(l,j) \\ &= \alpha_0^2 \sum_{j \in \mathbb{N}} b_j^2 + 2\alpha_0 \sum_{j=1}^n \alpha_j b_j + \sum_{j=1}^n \alpha_j^2 = \alpha_0^2 \sum_{j \in \mathbb{N}} b_j^2 + \sum_{j=1}^n \alpha_j (2\alpha_0 b_j + \alpha_j). \end{aligned}$$

If $\alpha_0 = 0$, then $B = \sum_{j=1}^n \alpha_j^2 > 0$ since we assumed $\alpha \neq 0$. Suppose $\alpha_0 \neq 0$. Then

$$B \geq \alpha_0^2 \sum_{j \in \mathbb{N}} b_j^2 + \sum_{j=1}^n \alpha_j^* (2\alpha_0 b_j + \alpha_j^*), \tag{18}$$

where

$$(\alpha_1^*,\ldots,\alpha_n^*) = \arg\min\left\{\sum_{j=1}^n \alpha_j(2\alpha_0 b_j+\alpha_j) : \sum_{j=1}^n \alpha_j = -\alpha_0\right\}. \tag{19}$$

Note that $(\alpha_1^*,\ldots,\alpha_n^*)$ is unique as the objective in (19) is strictly convex, which is minimized over a convex set. To solve (19), let us consider the Lagrangian, given as

$$L(\alpha_1,\ldots,\alpha_n,\lambda) = \sum_{j=1}^n \alpha_j(2\alpha_0 b_j+\alpha_j) - \lambda\Big(\sum_{j=1}^n \alpha_j+\alpha_0\Big),$$

where $\lambda \geq 0$. Differentiating L w.r.t. α_j and setting it to zero yields $\alpha_j^* = (\lambda - 2\alpha_0 b_j)/2$. Since $\sum_{j=1}^n \alpha_j^* = -\alpha_0$, we have $\lambda = \frac{2\alpha_0(a-1)}{n}$, where $a := \sum_{j=1}^n b_j$. Substituting for λ in α_j^*, we have

$$\alpha_j^* = \frac{\alpha_0(a-1-nb_j)}{n},\ j \in \mathbb{N}_n.$$

Substituting for α_j^* in (18) gives

$$B \geq \alpha_0^2 \sum_{j\in\mathbb{N}} b_j^2 + \frac{\alpha_0^2(a-1)^2}{n} - \alpha_0^2 \sum_{j=1}^n b_j^2 = \alpha_0^2 \sum_{j=n+1}^{\infty} b_j^2 + \frac{\alpha_0^2(\sum_{j=1}^n b_j - 1)^2}{n} > 0.$$

Consequently, we have $B > 0$ in any case, and therefore k is conditionally strictly pd. In the following, we however show that k is not characteristic.

Let $\mathbb{P} = \delta_0$ and $\mathbb{Q} = \sum_{j=1}^n b_j \delta_j$. Clearly $\mathbb{P} \neq \mathbb{Q}$. Consider

$$\begin{aligned}\Big\|\int_X k(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\Big\|_{\mathcal{H}}^2 &= \Big\|k(\cdot,0) - \sum_{j\in\mathbb{N}} k(\cdot,j)b_j\Big\|_{\mathcal{H}}^2 \\ &= k(0,0) - 2\sum_{j\in\mathbb{N}} k(j,0)b_j + \sum_{l,j\in\mathbb{N}} k(l,j)b_l b_j \\ &= \sum_{j\in\mathbb{N}} b_j^2 - 2\sum_{j\in\mathbb{N}} b_j^2 + \sum_{j\in\mathbb{N}} b_j^2 = 0.\end{aligned}$$

This implies the embedding in (2) is not injective and therefore k is not characteristic.

When X is finite, then the converse to Proposition 6 holds, that is, conditionally strictly pd kernels are characteristic, which is shown as follows. Let $X = \mathbb{N}_n$. Suppose k is conditionally strictly pd, that is, for any $n \geq 2$, $(\alpha_1,\ldots,\alpha_n) \neq (0,\ldots,0)$ with $\sum_{j=1}^n \alpha_j = 0$, and all distinct $x_1,\ldots,x_n \in X$, we have $\sum_{l,j=1}^n \alpha_l\alpha_j k(x_l,x_j) > 0$. Let $I := \{j : \alpha_j > 0\}$. Define $\mathbb{P} := \beta^{-1}\sum_{j\in I}\alpha_j\delta_j$ and $\mathbb{Q} := -\beta^{-1}\sum_{j\notin I}\alpha_j\delta_j$, where $\beta := \sum_{j\in I}\alpha_j$ and $\mathbb{P} \neq \mathbb{Q}$. Then

$$\Big\|\int k(\cdot,x)\,d(\mathbb{P}-\mathbb{Q})(x)\Big\|_{\mathcal{H}}^2 = \beta^{-2}\sum_{l,j=1}^n \alpha_l\alpha_j k(l,j) > 0$$

and therefore k is characteristic.

5. Conclusions

In this work, we have presented a unified study to explain the relation between universal kernels, characteristic kernels and RKHS embedding of measures: while characteristic kernels are related to the injective RKHS embedding of Borel probability measures, the universal kernels are related to the injective RKHS embedding of finite signed Borel measures. We showed that for all practical purposes (e.g., Gaussian kernel, Laplacian kernel, etc.), the notions of characteristic and universal kernels are equivalent. In addition, we also explored their relation to various other notions of positive definite (pd) kernels: strictly pd, integrally strictly pd and conditionally strictly pd. As an example, we showed all these notions to be equivalent (except for conditionally strictly pd) in the case of radial kernels on $\mathbb{R}^d$. We would like to note that while this study assumes the kernel to be real-valued, all the results extend verbatim to the case of complex-valued kernels as well.

This unified study shows that certain families of kernels, for example, bounded continuous translation invariant kernels on $\mathbb{R}^d$ and radial kernels on $\mathbb{R}^d$, are interesting for practical use, since the disparate notions of universal and characteristic kernels seem to coincide for these families. On the other hand, it may not give a guide regarding which kernel should be used given a problem.

Acknowledgments

The authors thank anonymous reviewers for their constructive comments that greatly improved the manuscript and also for pointing out to Suquet (2009). B. K. S. and G. R. G. L. wish to acknowledge support from the Institute of Statistical Mathematics (ISM), Tokyo, the National Science Foundation (grant DMS-MSPA 0625409), the Fair Isaac Corporation and the University of California MICRO program. Most of this work was done when B. K. S. was affiliated with the University of California, San Diego, of which a part was carried out while B. K. S. was visiting ISM. K. F. was supported by JSPS KAKENHI 19500249 and 22300098.

Appendix A. Radial Kernels are Translation Invariant on $\mathbb{R}^d$

Let k be radial on $\mathbb{R}^d \times \mathbb{R}^d$. Define $k(x,y) = \psi(x-y) := \int_{[0,\infty)} e^{-t\|x-y\|_2^2}\, d\nu(t), x,y \in \mathbb{R}^d$, where $\nu \in M_b^+([0,\infty))$. Since

$$e^{-t\|x-y\|_2^2} = \int_{\mathbb{R}^d} e^{-\sqrt{-1}(x-y)^T\omega}(4\pi t)^{-d/2} e^{-\|\omega\|_2^2/4t}\, d\omega,$$

we have $\psi(x) = \int_{\mathbb{R}^d} e^{-\sqrt{-1}x^T\omega}\phi(\omega)\, d\omega$, where

$$\phi(\omega) = \int_{[0,\infty)} (4\pi t)^{-d/2} e^{-\|\omega\|_2^2/4t}\, d\nu(t).$$

It is easy to check that $\phi(\omega) \geq 0, \forall\, \omega \in \mathbb{R}^d$ and $\phi \in L^1(\mathbb{R}^d)$. Therefore ψ satisfies (5), which means k is a bounded continuous translation invariant kernel on $\mathbb{R}^d$.

Appendix B. Supplementary Results

For completeness, we present the following supplementary result, which is a simple generalization of the technique used in the proof of Theorem 3 in Sriperumbudur et al. (2008).

Lemma 7 *Let k be a measurable and bounded kernel on a measurable space, X and let $\mathcal{H}$ be its associated RKHS. Then, for any $f \in \mathcal{H}$ and for any finite signed Borel measure, μ,*

$$\int_X f(x)\,d\mu(x) = \int_X \langle f, k(\cdot,x)\rangle_{\mathcal{H}}\,d\mu(x) = \left\langle f, \int_X k(\cdot,x)\,d\mu(x)\right\rangle_{\mathcal{H}}.$$

Proof Let $T_\mu : \mathcal{H} \to \mathbb{R}$ be a linear functional defined as $T_\mu[f] := \int_X f(x)\,d\mu(x)$. It is easy to show that

$$\|T_\mu\| := \sup_{f\in\mathcal{H}} \frac{|T_\mu[f]|}{\|f\|_{\mathcal{H}}} \leq \sqrt{\sup_{x\in X} k(x,x)}\|\mu\| < \infty.$$

Therefore, T_μ is a bounded linear functional on $\mathcal{H}$. By the Riesz representation theorem (Folland, 1999, Theorem 5.25), there exists a unique $\lambda_\mu \in \mathcal{H}$ such that $T_\mu[f] = \langle f, \lambda_\mu\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Set $f = k(\cdot,u)$ for some $u \in X$, which implies $\lambda_\mu = \int_X k(\cdot,x)\,d\mu(x)$ and the result follows. ■

The following result in Theorem 8 characterizes strictly pd kernels on $\mathbb{T}$, which we quote from Menegatto (1995). Before we state the result, we introduce some notation. For natural numbers m and n and a set A of integers, $m+nA := \{j \in \mathbb{Z} \mid j = m+na,\, a \in A\}$. An increasing sequence $\{c_l\}$ of nonnegative integers is said to be *prime* if it is not contained in any set of the form $p_1\mathbb{N} \cup p_2\mathbb{N} \cup \cdots \cup p_n\mathbb{N}$, where $p_1, p_2, \ldots, p_n$ are prime numbers. Any infinite increasing sequence of prime numbers is a trivial example of a prime sequence. We write $\mathbb{N}_n^0 := \{0,1,\ldots,n\}$.

Theorem 8 (Menegatto 1995) *Let ψ be a pd function on $\mathbb{T}$ of the form in (7). Let $\overline{N} := \{|n| : A_\psi(n) > 0,\, n \in \mathbb{Z}\} \subset \mathbb{N} \cup \{0\}$. Then ψ is strictly pd if $\overline{N}$ has a subset of the form $\cup_{l=0}^{\infty}(b_l + c_l\mathbb{N}_l^0)$, in which $\{b_l\} \cup \{c_l\} \subset \mathbb{N}$ and $\{c_l\}$ is a prime sequence.*

Appendix C. c_b-universality

As mentioned in Section 2, the definition of c_0-universality deals with $\mathcal{H}$ being dense in $C_0(X)$ w.r.t. the uniform norm, where X is an LCH space. Although the notion of c_0-universality addresses limitations associated with both *c*- and *cc*-universality, it only approximates a subset of $C(X)$, that is, it cannot deal with functions in $C(X)\backslash C_0(X)$. This limitation can be addressed by considering a larger class of functions to be approximated.

To this end, one can consider a notion of universality that is stronger than c_0-universality: a bounded continuous kernel, k is said to be c_b-universal if its corresponding RKHS, $\mathcal{H}$ is dense in $C_b(X)$, the space of bounded continuous functions on a topological space, X (note that $C_0(X) \subset C_b(X)$). This notion of c_b-universality may be more applicable in learning theory than c_0-universality as the target function, $f^\star$ can belong to $C_b(X)$ (which is a more natural assumption) instead of it being restrained to $C_0(X)$ (note that $C_0(X)$ only contains functions that vanish at infinity). Similar to Proposition 2, the following theorem provides a necessary and sufficient condition for k to be c_b-universal. Before we state the result, we need some definitions.

A *set function* is a function defined on a family of sets, and has values in $[-\infty,+\infty]$. A set function μ defined on a family τ of sets is said to be *finitely additive* if $\emptyset \in \tau$, $\mu(\emptyset) = 0$ and $\mu(\cup_{l=1}^n A_l) = \sum_{l=1}^n \mu(A_l)$, for every finite family $\{A_1,\ldots,A_n\}$ of disjoint subsets of τ such that $\cup_{l=1}^n A_l \in \tau$. A *field of subsets* of a set X is a non-empty family, Σ, of subsets of X such that $\emptyset \in \Sigma$, $X \in \Sigma$, and for all $A, B \in \Sigma$, we have $A \cup B \in \Sigma$ and $B\backslash A \in \Sigma$. An additive set function μ defined on a field Σ of subsets of a topological space X is said to be *regular* if for each $A \in \Sigma$ and $\varepsilon > 0$, there exists $B \in \Sigma$ whose closure is contained in A and there exists $C \in \Sigma$ whose interior contains A such that $|\mu(D)| < \varepsilon$ for every $D \in \Sigma$ with $D := C\backslash B$.

Proposition 9 (c_b-universality and RKHS embedding of set functions) *Suppose X is a normal topological space and $M_{rba}(X)$ is the space of all finitely additive, regular, bounded set functions defined on the field generated by the closed sets of X. Then, a bounded continuous kernel, k is c_b-universal if and only if the embedding,*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu,\ \mu \in M_{rba}(X), \tag{20}$$

is injective.

Proof The proof is very similar to that of Proposition 2, wherein we identify $(C_b(X))' \cong M_{rba}(X)$ such that $T \in (C_b(X))'$ and $\mu \in M_{rba}(X)$ satisfy $T(f) = \int_X f\, d\mu$, $f \in C_b(X)$ (Dunford and Schwartz, 1958, p. 262). Here, $\cong$ represents the isometric isomorphism. The rest of the proof is verbatim with $M_b(X)$ replaced by $M_{rba}(X)$. ■

Note that $M_{rba}(X)$ does not contain any measure—though a set function in $M_{rba}(X)$ can be extended to a measure—as measures are countably additive and defined on a σ-field. Since μ in Proposition 9 is not a measure but a finitely additive set function defined on a field, it is not clear how to deal with the integral in (20). Due to the technicalities involved in dealing with set functions, the analysis of c_b-universality and its relation to other notions considered in Section 3 is not clear, although it is an interesting problem to be resolved because of its applicability in learning theory.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Spring Verlag, New York, 1984.

A. Caponnetto, M. Pontil, C. Micchelli, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

W. Dahmen and C. A. Micchelli. Some remarks on ridge functions. *Approx. Theory Appl.*, 3: 139–143, 1987.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK, 2002.

N. Dunford and J. T. Schwartz. *Linear Operators. I: General Theory*. Wiley-Interscience, New York, 1958.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York, 1999.

K. Fukumizu, F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 473–480, 2009.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

E. Hewitt. Linear functionals on spaces of continuous functions. *Fundamenta Mathematicae*, 37: 161–189, 1950.

G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

V. A. Menegatto. Strictly positive definite kernels on the circle. *Rocky Mountain Journal of Mathematics*, 25(3):1149–1163, 1995.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

A. Pinkus. Strictly positive definite functions on a real inner product space. *Adv. Comput. Math.*, 20:263–271, 2004.

W. Rudin. *Functional Analysis*. McGraw-Hill, USA, 1991.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proc. of the 14th Annual Conference on Learning Theory*, pages 416–426, 2001.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, UK, 2004.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In R. Servedio and T. Zhang, editors, *Proc. of the 21st Annual Conference on Learning Theory*, pages 111–122, 2008.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press, 2009.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *JMLR Workshop and Conference Proceedings*, volume 9, pages 781–788. AISTATS, 2010a.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010b.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

J. Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain Journal of Mathematics*, 6(3):409–433, 1976.

Ch. Suquet. Reproducing kernel Hilbert spaces and random measures. In H. G. W. Begehr and F. Nicolosi, editors, *Proc. of the 5th International ISAAC Congress, Catania, Italy, 25-30 July 2005*, pages 143–152. World Scientific, 2009.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.